Environmental and Resource Economics **20:** 241–254, 2001.

What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?

Evidence from a Meta-Analysis

JOHN A. LIST[1] and CRAIG A. GALLET[2,*]
[1] *University of Maryland, College, Park, MD 85721, USA;* [2] *California State University at Sacramento, Sacramento, CA 95819-6016, USA*
(** Corresponding author: E-mail: cgallet@juno.com*)

Accepted 18 April 2001

Abstract. Preferences elicited in hypothetical settings have recently come under scrutiny, causing estimates from the contingent valuation method to be challenged due to perceived "hypothetical bias." Given that the received literature derives value estimates using heterogeneous experimental techniques, understanding the effects of important design parameters on the magnitude of hypothetical bias is invaluable. In this paper, we address this issue statistically by using a meta-analysis to examine data from 29 experimental studies. Our empirical findings suggest that on average subjects overstate their preferences by a factor of about 3 in hypothetical settings, and that the degree of over-revelation is influenced by the distinction between willingness-to-pay and willingness-to-accept, public versus private goods, and several elicitation methods.

Key words: CVM, hypothetical bias, meta-analysis

JEL classification: Q26, Q28, H41

1. Introduction

Understanding why people misstate their actual preferences for a good when asked a hypothetical question remains an important issue in nonmarket valuation. While biases have been observed in both directions, much work in this literature suggests that people tend to overstate their actual willingness to pay in hypothetical situations.[1] In response, the National Oceanic and Atmospheric Administration's (NOAA) blue-ribbon panel composed of hall-of-fame economists, such as Kenneth Arrow and Robert Solow, recommended that hypothetical bids be deflated using a "divide by 2" rule unless these bids can be calibrated using actual market data (NOAA 1994, 1996). The NOAA rule has triggered a search for a calibration function to correct systematic bias between intentions and actions in valuation exercises (e.g., Blackburn et al. 1994; Hofler and List 2000). Although a fair amount of literature has resulted, the calibration procedure is not universally accepted. For example, one calibration critic eloquently states that: "[t]he calibration issue, it seems to me, is an audacious attempt to promote a Kuhnian paradigm shift ... I would argue vigorously that the essential premise is unproven and the question is

therefore premature and presumptuous. The proposed new calibration paradigm is at this moment merely a rambunctious challenger to the dominant external validation paradigm" (Randall 1996, p. 200). We interpret this statement as a call for a more thorough examination of the technical aspects that influence the reported calibration functions in the literature.

In this paper, we make this next step not by running a new field or laboratory experiment, but by using a meta-analysis. A primary advantage of using a meta-analysis is that it allows us to take a step back from the burgeoning literature to determine whether important experimental parameters systematically affect the relationship between hypothetical and actual responses. While this statistical approach has been used in the past to uncover relationships ranging from the proper determinants of gasoline demand (Espey 1998) to estimating the effects of environmental regulations on new firm location patterns (Jeppessen et al. 2001), it is particularly useful for a controversial issue such as comparing hypothetical and actual reported valuations since a wide array of methodologies has been used to gather data.

Our data set is comprised of 174 observations collected from 29 experimental studies. The data have quite a broad range: from willingness to pay estimates for a Cal Ripken Jr. baseball card at a sportscard show to compensation demanded for a Wisconsin goose license. Our primary line of inquiry will provide evidence pertaining to the effects of various experimental protocol on the observed calibration factors. For example, amongst other important issues, we provide insights into the following questions: 1) Does hypothetical bias exist in the typical contingent valuation exercise? If so, what is the magnitude of the bias? 2) Does the distinction between willingness-to-pay (WTP) and willingness-to-accept (WTA) measures of value significantly influence the hypothetical/actual ratio? 3) Do various elicitation methods, such as dichotomous choice, Vickrey 2nd price auction, or random nth price auction, affect the calibration factor? 4) Do within-subject experiments provide larger calibration factors than between-subject experiments? 5) Do laboratory and field experiments yield similar calibration functions? 6) Do public goods tend to have larger calibration factors than private goods?

The remainder of our study is crafted as follows. The next section sets the stage with a broad review of the calibration literature. Section 3 presents out reduced-form empirical model, while the estimation results are discussed in Section 4. The paper concludes with a summary in Section 5.

2. Previous Work

In an effort to include as many studies as possible, we made the pragmatic choice to focus on studies that explicitly include discussion of experimental design variables that are commonly believed to affect stated preferences. This procedure allows us to include a significant amount of the literature in our meta-analysis without compromising significantly the character of the empirical model. Although

our review is broad, a significant growth in the literature in the past few years makes our attempt to provide a summary of the studies as merely representative of the overall body of knowledge.[2] Our final tally includes 29 studies that provide 174 total observations across both hypothetical and actual valuations.[3] Important methodological features across these studies include the setting of the experiment (laboratory, field, or both), type of good (public or private), type of comparison (within- or between-person study), and elicitation method (open-ended, Vickrey 2nd price auction, dichotomous choice, first price sealed bid auction, random nth price auction, provision point mechanism, Smith auction, or Becker, Degroot, and Marschak (BDM)).

Table I provides an overview of the features of each study, as well as the calibration factors (mean hypothetical divided by mean actual) reported in each study. Before proceeding, it is worthwhile to mention a few important characteristics inherent in our study. First, we refer to "hypothetical bias" as the difference between hypothetical and actual statements of value, where actual statements of value are obtained from experiments with real economic commitments. As such, we are assuming that the cash-based estimates are unbiased. While most of the studies in the literature use incentive compatible mechanisms to obtain value estimates, some studies use non-incentive compatible institutions (e.g., an open-ended survey), perhaps leading our baseline (cash-based estimates) to be biased. To control for this potential nuance, we include regressors in our empirical model that allow a robust comparison within and across elicitation mechanisms. Second, note that for several of the studies a range of calibration factors is provided because the authors report more than one estimate. In these cases, for efficiency purposes, we make use of the entire spectrum of information by running different sets of regressions, as described below.

The top panel of Table I shows a chronological ordering of WTP studies that report hypothetical and actual statements. The research appears to have commenced with Bohm's (1972) seminal experimental lab study which compared bids in hypothetical and actual experimental markets that elicited subjects' stated value to sneak preview a Swedish television show. Given that reported calibration factors tend to exceed 1, Bohm's (1972) results suggest that people moderately overstate their actual values when asked a hypothetical question. Subsequent lab research has generally supported Bohm's findings (e.g., Bishop and Heberlein 1979; Neill et al. 1994; Fox et al. 1998; List and Shogren 1998a, b; Balistreri et al. 1998).

Exceptions to this upward bias can be found in numerous studies (e.g., Sinden 1988; Johannesson et al. 1998), but the average person seems to exaggerate his or her actual WTP across a broad spectrum of goods with vastly different experimental protocol. For instance, average hypothetical bids for baseball cards were nearly 3.5 times larger than associated actual bids, which is in the range of calibration factors observed for irradiated/non-irradiated pork and water color paintings and maps (see List and Shogren 1998a; Fox et al. 1998; Neill et al. 1994). Overall,

Table I. Summary of studies.

Study	Year	Type of experiment	Type of good	Type of comparison	Type of elicitation	Calibration factor[a]
Willingness-to-pay:						
Bohm	1972	Laboratory	Private	Within group	Open-ended	1.00
						1.16
						1.16
						1.34
Bishop and Heberlein	1979	Field	Private	Between group	Dichotomous	0.30–1.60
Bishop and Heberlein	1986	Field	Private	Between group	Open-ended	1.30–2.30
					Dichotomous	0.80
Brookshire and Coursey	1987	Field and lab	Public	Between group	Smith	2.00
						1.85
Dickie et al.	1987	Field	Private	Between group	Dichotomous	1.00
Coursey et al.	1987	Laboratory	Private	Between group	Vickrey	1.00
Sinden et al.	1988	Laboratory	Public	Within group	Open-ended	0.80–1.50
Kealy et al.	1988	Laboratory	Private	Between group	Dichotomous	1.40
			Public	Within group		1.00–2.00
Seip and Strand	1992	Field	Public	Within group	Dichotomous	10.30
Navrud	1992	Field	Private	Within group	Dichotomous	3.20
			Public			1.60–2.10
Boyce et al.	1992	Field and lab	Private	Between group	BDM	1.50
						2.10
						0.90
McClelland et al.	1993	Laboratory	Private	Between group	Vickrey	2.20
						0.80
Neill et al.	1994	Laboratory	Private	Between group	Open-ended	3.1–25.1
Loomis et al.	1996	Laboratory	Private	Between group	Open-ended	1.80–2.90
				Within group	Dichotomous	2.00–3.60
Brown et al.	1996	Field	Public	Between group	Open-ended	4.10
					Dichotomous	6.50
Frykblom	1997	Laboratory	Private	Between group	Dichotomous	1.50
Loomis et al.	1997	Laboratory	Private	Between group	Open-ended	1.86
					Dichotomous	2.54–3.00
Spencer et al.	1998	Laboratory	Public	Between group	provision pt.	4.67
						4.66
List and Shogren	1998a	Field	Private	Within group	Vickrey	2.54
						3.47
						2.19
Fox et al.	1998	Laboratory	Private	Between group	Vickrey	1.20
				Within group		1.50
Balistreri et al.	1998	Laboratory	Private	Between group	open-ended	1.25
					Dichotomous	1.54
					Dichotomous	0.58

Table I. Continued.

Study	Year	Type of experiment	Type of good	Type of comparison	Type of elicitation	Calibration factor[a]
Johannesson et al.	1998	Laboratory	Private	Between group	Dichotomous	1.18
				Between group		0.80
				Within group		1.29
				Within group		0.88
Frykblom	2000	Laboratory	Private	Between group	Vickrey	1.33
Willingess-to-accept:						
Bishop et al.	1983	Field	Private	Between group	Dichotomous	1.60
Coursey et al.	1987	Laboratory	Private	Between group	Vickrey	2.00
Brookshire and Coursey	1987	Field and lab	Public	Between group	Smith	28.20
						25.79
Bishop and Heberlein	1990	Field	Private	Between group	Sealed bid	0.70
					Dichotomous	2.74
Smith and Mansfield	1998	Field	Private	Between group	Dichotomous	1.00
List and Shogren	1998b	Laboratory	Private	Within group	Random price	1.42
List and Shogren	1999	Laboratory	Private	Within group	Random price	0.70–1.66

[a]Calibration factor is calculated as mean hypothetical/mean actual.

the calibration factors are in the range of calibration factors, 1.0 to 10.0, observed in earlier work (Diamond and Hausman 1994), and reinforce the argument that people tend to overstate their actual WTP when confronted with a hypothetical valuation question.

Interestingly, researchers have spent much less energy on understanding the relationship between real and hypothetical compensation demanded (WTA) measures of value. The bottom panel of Table I reveals that empirical evidence from this relatively small lot of studies is mixed. List and Shogren (1999) calibrate real and hypothetical WTA estimates elicited for consumer goods in a multi-unit, random nth-price auction. Using a within-subject experimental design, they find that people understated their real willingness to accept in the hypothetical regimes. Bishop and colleagues found that Wisconsin goose hunters' overstated their actual WTA to sell goose-licenses, whereas deer hunters' in a sealed-bid auction understated their actual WTA to sell deer-permits. Smith and Mansfield's (1998) field experimental results suggest that real and hypothetical WTA statements for the opportunity to spend time in a second set of interviews on an undisclosed topic are statistically equivalent. As one can readily see, experimental results from the WTA literature are literally all over the map.

The underlying story in Table I is that hypothetical bias appears to exist in contingent valuation exercises across a broad spectrum of goods. Furthermore, the raw data in Table I imply that the relationship between real and hypothetical stated values may be specific to important experimental protocol, and that attempts to bridge the gap statistically might be futile unless we understand the experimental

factors that cause these discrepancies. Yet, one important issue that perusal of Table I cannot resolve is the extent and manner in which each of the individual experimental design parameters influences the actual/hypothetical relationship.

3. Empirical Model

To obtain information on the important factors that influence the magnitude of reported hypothetical bias, we estimate the following reduced-form model:

$$CF = X'\beta + u, \tag{1}$$

where CF is the natural logarithm of the calibration factor (mean hypothetical value divided by the mean actual value); X is a nonstochastic n by m matrix with rank m, where the m regressors are experimental design variables that are common to the broad range of studies.[4] Regressors in X include: $X_1 = 1$ if laboratory is strictly used, 0 otherwise;[5] $X_2 = 1$ if WTP study, 0 if WTA study; $X_3 = 1$ if good is private, 0 if good is public; $X_4 = 1$ if comparison is within-group, 0 if comparison is between group; $X_5 = 1$ if elicitation method is open-ended, 0 otherwise; $X_6 = 1$ if elicitation method is first price sealed bid, 0 otherwise; $X_7 = 1$ if elicitation method is provision point mechanism, 0 otherwise; $X_8 = 1$ if elicitation method is Smith auction, 0 otherwise; $X_9 = 1$ if elicitation method is random nth price auction, 0 otherwise; $X_{10} = 1$ if elicitation method is BDM, 0 otherwise; and $X_{11} = 1$ if elicitation method is dichotomous choice, 0 otherwise. μ is the well-behaved error team.

A few aspects of equation (1) merit further consideration. First, given that several studies report a range for the calibration factor, we use three different regressand constructs, coinciding with the minimum, median, and maximum values of the calibration factor reported. Interestingly, sample means within these three categories are 2.79, 3.05, and 3.31. In general, these sample statistics suggest that important discrepancies exist across hypothetical and actual statements – on average subjects overstate their preferences by a factor of about 3 in hypothetical exercises. In addition, the discrepancy between the minimum and maximum reported calibration factors is relatively small. Second, setting each of the dichotomous regressors to zero yields the base-case scenario. Accordingly, as a point of reference, the base case consists of a between-sample (field or lab and field) WTA study of a public good that uses a Vickrey 2nd price auction to elicit individual valuations. Third, other regressors could have also been included in X, but given that our focus pertains to important experimental protocol, we chose to examine the design parameters reported in the majority of published studies. This procedure allows a much broader analysis while not significantly compromising the generality of our results.

Table II. Empirical results.[a]

Variable	Estimated coefficients		
	Minimum	Median	Maximum
Laboratory (X_1)	−0.27 (0.22)	−0.32 (0.23)	−0.34 (0.24)
Willingness-to-pay (X_2)	−0.75 (0.32)**	−0.65 (0.33)**	−0.61 (0.35)*
Private good (X_3)	−0.62 (0.29)**	−0.64 (0.30)**	−0.68 (0.32)**
Within group (X_4)	0.08 (0.21)	−0.01 (0.22)	−0.04 (0.23)
Type of elicitation:			
Open-ended (X_5)	−0.09 (0.27)	0.15 (0.28)	0.28 (0.29)
First price sealed bid (X_6)	−1.71 (0.73)**	−1.70 (0.75)**	−1.69 (0.79)**
Provision point (X_7)	0.58 (0.59)	0.54 (0.61)	0.49 (0.64)
Smith auction (X_8)	0.38 (0.51)	0.32 (0.53)	0.28 (0.56)
Random price auction (X_9)	−1.18 (0.61)*	−0.76 (0.63)	−0.52 (0.66)
Becker, Degroot, and Marschak (X_{10})	−0.26 (0.45)	−0.34 (0.47)	−0.37 (0.50)
Dichotomous choice (X_{11})	−0.36 (0.24)	−0.30 (0.25)	−0.26 (0.26)
Constant	1.97 (0.47)**	1.98 (0.49)**	2.01 (0.51)***
Sample size	58	58	58
R-squared	0.50	0.46	0.43
F	4.10	3.56	3.17

[a]The natural log of calibration factor is the dependent variable. Given that a range is often reported for the calibration factor, results for three constructions (using the minimum, median, and maximum values) of the dependent variable are reported. Standard errors are provided in parentheses to the right of the estimated coefficients. The F-statistic for testing the significance of the overall model is provided at the bottom of the table.
*Significant at the 10% level.
**Significant at the 5% level.
***Significant at the 1% level.

4. Estimation Results

Estimation results of equation (1), which are reported in Table II, show that all three of the estimated model types are significant at conventional levels via an *F*-test ($F = 4.1$, 3.56, and 3.17 for the minimum, median, and maximum models).[6] The minimum and median empirical specifications explain 50 percent and 46 percent of the variation in the regressand, whereas the maximum model type explains 43 percent of the variation.

Several interesting results emerge from an examination of the individual estimated coefficients. First, although each of the estimated models as a whole is significant, significance of the individual estimates coefficients is rare. Yet, one result that is robust across the three models is the significance of the coefficient estimates of the dichotomous regressors for willingness-to-pay and private good. For the former regressor, the coefficient estimate is consistently negative,

suggesting that, *ceteris paribus*, the calibration factor obtained from a WTP study will be lower than a comparable calibration factor from a WTA study. This makes sense, as most subjects should be more apt to correctly state their true preferences when performing a familiar hypothetical task (WTP) rather than an unfamiliar one (WTA) (see, e.g., Cummings et al. 1986). Similar reasoning can be used to explain the negative coefficient estimate associated with private goods. Since most subjects are more comfortable valuing goods they commonly purchase, they may make less errors in valuing these types of goods than valuing public goods, which they may have little valuation experience.[7]

Referring to the elicitation method variables, based on F-tests of the joint significance of the elicitation variables, we find that elicitation technique matters. In particular, except for the median model, the F-statistics pertaining to elicitation are significant at the $p < 0.10$ level (i.e., $F = 2.10$ (minimum), $F = 1.70$ (median), and $F = 2.40$ (maximum)). Furthermore, with respect to the individual effects, comparing elicitation methods yields interesting results. However, since the significance of the estimated coefficients of the elicitation method variables is interpreted relative to the baseline, one needs to be careful with such interpretation. Accordingly, to gauge pair-wise differences in calibration factors across all elicitations, while controlling for modeling technique, we perform a series of *t*-tests allowing for different baseline comparisons. Empirical results indicate that calibration factors differ across several elicitation methods. For example, the following results hold for the minimum model: (i) calibration factors obtained Vickrey 2nd price auctions (current baseline) are greater than calibration factors from from random nth price auctions ($t = -1.94$) and first price sealed bid auctions ($t = -2.36$); (ii) first-price sealed bid auctions yield lower calibration factors than open-ended elicitation schemes ($t = -2.19$), provision point mechanisms ($t = -2.55$), dichotomous choice institutions ($t = -1.91$), Smith auctions ($t = -2.68$), and the BDM method ($t = -1.81$); (iii) random nth price auctions yield lower calibration factors than open-ended mechanisms ($t = -1.79$), provision point mechanisms ($t = -2.15$), and Smith auctions ($t = -2.12$); and (iv) calibration factors for dichotomous choice questions are lower than factors from the provision point mechanism ($t = -1.72$) and Smith auctions ($t = -1.66$).[8]

While the remaining individual coefficient estimates are not significantly different from zero at conventional levels, they do provide interesting insights. For example, the data suggest that the calibration factor is not affected by whether the experiment takes place in the lab or field. This finding is encouraging since it provides evidence that nuances such as subject pools, social distance, and subtleties associated with the laboratory setting may not compromise the generality of the empirical findings. In addition, there has been some discussion pertaining to the use of between-subject versus within-subject experimental designs (see, e.g., Cummings et al. 1995). We find that calibration factors are not significantly different between the two treatment types, and therefore we tentatively recommend use of within-sample experiments. We have not found any prior evidence

Table III. Empirical results.[3]

Variable	Estimated coefficients		
	Minimum	Median	Maximum
Laboratory (X_1)	−0.39 (0.19)*	−0.28 (0.22)	−0.30 (0.23)
Willingness-to-pay (X_2)	−0.51 (0.28)*	−0.56 (0.32)*	−0.53 (0.33)
Private good (X_3)	−0.48 (0.26)*	−0.56 (0.29)*	−0.61 (0.31)*
Within group (X_4)	−0.04 (0.18)	−0.06 (0.21)	−0.09 (0.22)
Type of elicitation:			
Open-ended (X_5)	−0.06 (0.23)	0.12 (0.27)	0.25 (0.28)
First price sealed bid (X_6)	−0.95 (0.64)	−0.93 (0.72)	−0.91 (0.76)
Provision point (X_7)	0.65 (0.52)	0.54 (0.59)	0.49 (0.62)
Smith auction (X_8)	0.44 (0.45)	0.41 (0.51)	0.36 (0.54)
Random price auction (X_9)	−0.53 (0.53)	−0.68 (0.60)	−0.46 (0.64)
Becker, Degroot, and Marschak (X_{10})	−0.38 (0.40)	−0.31 (0.45)	−0.33 (0.48)
Dichotomous choice (X_{11})	−0.18 (0.21)	−0.20 (0.24)	−0.17 (0.25)
Constant	1.79 (0.41)***	1.84 (0.47)***	1.88 (0.49)***
Sample size	58	58	58
R-squared	0.50	0.43	0.40
F	4.15	3.11	2.74

[a]The absolute value of the natural log of the calibration factor is the dependent variable. Given that a range is often reported for the calibration factor, results for three constructions (using the minimum, median, and maximum values) of the dependent variable are reported. Standard errors are provided in parentheses to the right of the estimated coefficients. The *F*-statistic for testing the significance of the overall model is provided at the bottom of the table.

suggesting that between-sample procedures are preferred, and therefore we believe that within-sample designs are appropriate because they have a unique advantage in that the researcher can control for potentially important individual-specific effects in the statistical analysis.

Although most of the reported calibration factors in Table I exceed one, several are less than one. If the goal is to gauge the degree of hypothetical bias, whether it is positive or negative, interpreting the results in Table II can be cumbersome. For example, depending on the calibration factor's relationship to one (no bias), taking the natural logarithm will result in both negative and positive regressand values. Accordingly, a negative coefficient estimate of a regressor can be interpreted as a factor causing a movement of the calibration factor towards or away from the point of zero bias. To control for this nuance, in Table III we report results from the estimation of equation (1) where the dependent variable is the absolute value of the natural log of the calibration factor. A positive (negative) coefficient estimate, therefore, implies that the respective regressor causes a greater (lesser) departure of the calibration factor from unity, or the no bias point.

Results in Table III again suggest there is some evidence that the disparity between hypothetical and actual statements is a function of whether the respondent is providing WTA or WTP values. In both the minimum and median empirical models, negative coefficient estimates, which are both significant at the $p < 0.10$ level, imply that responses in WTP settings tend to correspond with actual WTP values more closely then hypothetical WTA values track actual WTA stated values. Empirical results measuring differences across public and private goods also are in accord with the results in Table II: hypothetical bias is considerably less for private goods compared to public goods. This result is consonant with the intuition provided above concerning the valuation of goods and services within familiar contexts.

Considering elicitation mechanisms, although the disparity is not sensitive to elicitation methods relative to the base case of Vickrey 2nd price auctions, when other baselines are used we do find significant differences in elicitation. For example, with respect to the minimum model, we find differences between the degree of bias across provision point mechanisms and dichotomous choice questions, values obtained from the BDM approach, random nth price auctions, and sealed bid auctions. Also, sealed bid auctions and Smith auctions yield significantly different calibration factors.

We should also note that we also estimated all of the models in Table II and III using a linear, rather than a logarithmic, regressand. These empirical estimates, which are available upon request, are never qualitatively different from parameter estimates presented in Tables II and III.[9] Yet the linear models tend to perform better than the logarithmic models in terms of: i) overall fit of the regression model and ii) joint significance of the elicitation variables. As such, in summarizing the empirical results, we believe that overall they suggest: (i) use of laboratory or field experiments do not systematically affect calibration factors (although the coefficient of the lab dummy variable is slightly significant in the minimum model of Table III); (ii) provide goods yield lower calibration factors than public goods; (iii) many of the theoretically incentive-compatible elicitation techniques do affect the calibration factor, suggesting that some methods induce more truthful responses than others; (iv) in terms of truthful revelation, our general results tend to support recommendations from CVM experts, who argue that WTP rather than WTA is the preferred valuation procedure (see, e.g., Cummings et al. 1986). Finally, in light of the most commonly used techniques from Table I (i.e., laboratory, WTP, private good, between group, and dichotomous), using the results in Table II the predicted calibration factors for the most prevalent type of study are 1.26 (minimum), 1.28 (median), and 1.30 (maximum), which suggests that the most common type of WTP study will tend to produce a slightly (upward) biased estimate of the actual value when obtaining a hypothetical statement of value.

5. Concluding Comments

Given that nonmarket valuation remains one of the most controversial issues in environmental economics, understanding the factors that cause disparities between hypothetical and actual reported valuations is invaluable. This paper provides a review of the laboratory and field evidence on the gap between intentions and actions and examines whether a systematic statistical relationship exists between words and deeds. Evidence from our meta-analysis suggests that certain experimental protocol influence deviations in hypothetical and actual statements. For example, willingness to pay studies yield smaller hypothetical-to-actual ratios than willingness to accept studies. In addition, we find that certain elicitation methods induce disparities between hypothetical and actual statements.

More research is necessary. Undeniably, our results should only be considered a first attempt at quantifying the various experimental methods that may affect hypothetical bias. Given the small sample size for certain elicitation techniques, we view it inappropriate to debate the merits of the various elicitation methods based on our results alone. At this early stage in the debate, we are comfortable with arguing that our results suggest that experimental procedures affect reported calibration functions in meaningful ways, and any calibration exercise that seeks reliability will need to understand the important experimental protocol which induce biases. We hope that our empirical results will lead to discoveries of more robust calibration functions.

Acknowledgement

We are grateful for the helpful comments of Mark Dickie, Glenn Harrison, and John Loomis. Any errors or omissions are the responsibility of the authors.

Notes

1. See, for example, Bohm (1972), Bishop and Heberlein (1979), Neill et al. (1994), Fox et al. (1998), and List and Shogren (1998a).
2. Since our focus is on economic comparisons between hypothetical and actual behavior within the context of valuation, we avoid discussion of other bodies of literature that compare behavior across hypothetical/actual regimes, such as risk-taking studies. Also, given that our main focus pertains to mean calibration factors (i.e., mean hypothetical/mean actual), studies that do not report both estimates are excluded (e.g., Cummings et al. 1995, 1997). We also do not include recent "cheap talk" studies (e.g., List 2001).
3. The literature review extends List and Shogren (1999) and Foster et al. (1997).
4. Regressions were performed with the non-transformed calibration factor as the dependent variable, with the signs of the estimates coefficients similar to those reported for the semi-log version.
5. Since studies vary according to whether results are obtained in a lab, field, or a combination of lab and field, ideally we would like to have two dummy variables in the model to account for these three possibilities. Unfortunately, this treatment results in a violation of the rank condition. Accordingly, to bypass this problem, we let X_1 equal one for studies that obtain calibration

factors strictly in the lab, while X_1 equal to zero refers to studies that use a field or a field and lab combination.

6. Each of the three regression models were examined for heteroskedasticity using a White Test. Due to the relatively small χ^2 values (i.e., $\chi^2 = 20.30$ (minimum), $\chi^2 = 8.7$ (median), and $\chi^2 = 7.54$ (maximum)), we do not correct for heteroskedasticity.
7. As further evidence, observations across the three formulations of the dependent variable were combined, such that the model was estimated via pooled OLS. Empirical results are not qualitatively different from those in Table II; the calibration factor remains sensitive to WTA versus WTP measures of value and private versus public good distinctions.
8. With respect to the median and maximum models, the sealed bid coefficient differs from the coefficients of open-ended, provision point, Smith, Vickrey 2nd price auction, and dichotomous choice. Also, open-ended differs from dichotomous choice.
9. Following a reviewer's suggestion, we also estimated the model with year of publication included as a regressor. This may control for the possibility that techniques are improving over time, thereby reducing the severity of hypothetical bias. With year of publication included, the direction and significance of the coefficients in Tables II and III do not change. Also, the results show that the magnitude of hypothetical bias is larger in later studies as compared to earlier studies. However, since most of the studies in Table I occur within a relatively short publication window (i.e., the last 15 years), coupled with the fact that the lag between submission date and publication date varies across journals, we did not include publication date in the reported specification. Yet the results are available from the author's upon request.

References

Balistreri, E., G. McClelland, G. Poe and W. Schulze (1998), *Can Hypothetical Questions Reveal True Values? A Laboratory Comparison of Dichotomous Choice and Open-Ended Contingent Values with Auction Values*. Cornell University, working paper, WP 97-15.

Bishop, R. and T. Heberlein (1979), 'Measuring Values of Extramarket Goods: Are Indirect Measures Biased?', *American Journal of Agricultural Economics* **61**, 926–930.

Bishop, R. (1986), 'Assessing the Validity of Contingent Valuations: Three Field Experiments', *Science of the Total Environment* **56**, 434–479.

Bishop, R. (1990), 'The Contingent Valuation Method', in R. L. Johnson and G. V. Johnson, eds., *Economic Valuation of Natural Resources*. Boulder, CO: Westview Press, pp. 81–104.

Bishop, R., T. Heberlein and M. J. Kealy (1983), 'Contingent Valuation of Environmental Assets: Comparisons with a Simulated Market', *Natural Resources Journal* **23**, 619–633.

Blackburn, M., G. Harrison and E. E. Rutström (1994), 'Statistical Bias Functions and Informative Hypothetical Surveys', *American Journal of Agricultural Economics* **76**, 1084–1088.

Bohm, P. (1972), 'Estimating Demand for Public Goods: An Experiment', *Europeawn Economic Review* **3**(2), 111–130.

Boyce, R., G. McClelland, T. Brown, G. Peterson and W. Schulze (1992), 'An Experimental Examination of Intrinsic Values as a Source of the WTA-WTP Disparity', *American Economic Review* **82**(5), 1366–1373.

Brookshire, D. and D. Coursey (1987), 'Measuring the Value of a Public Good: An Empirical Comparison of Elicitation Procedures', *American Economic Review* 77(4): 554–566.

Brown, T., P. Champ, R. Bishop and D. McCollum (1996), 'Which Response Format Reveals the Truth about Donations to a Public Good?', *Land Economics* **72**(2), 152–166.

Coursey, D., J. Hovis and W. Schulze (1987), 'The Disparity between Willingness to Accept and Willingness to Pay Measures of Value', *Quarterly Journal of Economics* **102**, 679–690.

Cummings, R., D. Brookshire and W. Schulze, eds. (1986), *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Totowa, NJ: Rowman and Allanheld.

Cummings, R., G. Harrison and E. E. Rutström (1995), 'Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive Compatible?', *American Economic Review* **85**, 260–266.

Cummings, R., S. Elliot, G. Harrison and J. Murphy (1997), 'Are Hypothetical Referenda Incentive Compatible?', *Journal of Political Economy* **105**(3), 609–621.

Diamond, P. and J. Hausman (1994), 'Contingent Valuation: Is Some Number Better than No Number?', *Journal of Economic Perspectives* **8**, 45–64.

Dickie, M., A. Fisher and S. Gerking (1987), 'Market Transactions and Hypothetical Demand Data: A Comparative Study', *Journal of American Statistical Association* **82**, 69–75.

Espey, M. (1998), 'Gasoline Demand Revisited: An International Meta-Analysis of Elasticities', *Energy Economics* **20**, 273–295.

Foster, V., I. Bateman and D. Harley (1997), 'Real and Hypothetical Willingness to Pay for Environmental Preservation: A Non-Experimental Comparison', *Journal of Agricultural Economics* **48**(2), 123–138.

Fox, J., J. Shogren; D. Hayes and J. Kliebenstein (1998), 'CVM-X: Calibrating Contingent Values with Experimental Auction Markets', *American Journal of Agricultural Economics* **80**, 455–465.

Frykblom, P. (1997), 'Hypothetical Question Modes and Real Willingness to Pay', *Journal of Environmental Economics and Management* **34**, 275–287.

Frykblom, P. (2000), 'Willingness to Pay and the Choice of Question Format: Experimental Results', *Applied Economics Letters* **7**, 665–667.

Grether, D. and C. Plott (1979), 'Economic Theory of Choice and the Preference Reversal Phenomenon', *American Economic Review* **69**(1), 623–638.

Hofler, R. and J. A. List (2000), *Valuation on the Frontier: Calibrating Actual and Hypothetical Statements*. University of Arizona, working paper.

Irwin, J., G. McClelland and W. Schulze (1992), 'Hypothetical and Real Consequences in Experimental Auctions for Insurance Against Low Probability Risks', *Journal of Behavioral Decision Making* **5**, 107–116.

Jeppessen, T., J. A. List and H. Folmer (2001), 'Environmental Regulations and New Plant Location Decisions: Evidence from a Meta-Analysis', *Journal of Regional Science*, forthcoming.

Johannesson, M., B. Liljas and P. O. Johansson (1998), 'An Experimental Comparison of Dichotomous Choice Contingent Valuation Questions and Real Purchase Decisions', *Applied Economics* **30**, 643–647.

Kealy, M., J. Dovidio and M. Rockel (1988), 'Accuracy in Valuation is a Matter of Degree', *Land Economics* **64**, 158–171.

Kealy, M., J. Montgomery and J. Dovidio (1990), 'Reliability and Predictive Validity of Contingent Values: Does the Nature of the Good Matter?', *Journal of Environmental Economics and Management* **19**, 244–263.

List, J. A. (2001), 'Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards', *American Economic Review*, forthcoming.

List, J. A. and J. Shogren (1998a), 'Calibration of the Difference between Actual and Hypothetical Valuations in a Field Experiment', *Journal of Economic Behavior and Organization* **37**(2), 193–205.

List, J. A. and J. Shogren (1998b), 'The Deadweight Loss of Christmas: Comment', *American Economic Review* **88**(5), 1350–1355.

List, J. A. and J. Shogren (1999), 'Calibration of Willingness-to-Accept', *Journal of Environmental Economics and Management*, forthcoming.

Loomis, J., T. Brown, T. Lucero and G. Peterson (1996), 'Improving Validity Experiments of Contingent Valuation Methods: Results of Efforts to Reduce the Disparity of Hypothetical and Actual Willingness to Pay', *Land Economics* **72**(4), 450–461.

Loomis, J., T. Brown, T. Lucero and G. Peterson (1997), 'Evaluating the Validity of the Dichotomous Choice Question Format in Contingent Valuation', *Environmental and Resources Economics* **10**, 109–123.

McClelland, G., W. Schulze and D. Coursey (1993), 'Insurance for Low-Probability Hazards: A Biomodal Response to Unlikely Events', *Journal of Risk and Uncertainty* **7**, 95–116.

Navrud, S. (1992), 'Willingness to Pay for Preservation of Species – An Experiment with Actual Payments', in S. Navrud, ed., *Pricing the European Environment*. Oslo: Scandinavian University Press; distributed by Oxford University Press, New York, pp. 231–246.

Neill, H., R. Cummings, P. Ganderton, G. Harrison and T. McGuckin (1994), 'Hypothetical Surveys and Real Economic Commitments', *Land Economics* **70**(2): 145–154.

National Oceanic and Atmospheric Administration (1994), 'Natural Resource Damage Assessment: Proposed Rules', *Federal Register*, 4 May **59**, 23098–23111.

National Oceanic and Atmospheric Administration (1996), 'Natural Resource Damage Assessment: Final Rules', *Federal Register*, 5 January **61**, 439.

Randall, A. (1996), 'Calibration of CV Responses: Discussion', in D. Bjornstad and J. Kahn, eds., *The Contingent Valuation of Environmental Resources*. London: Edgar Elgar, pp. 198–207.

Seip, K. and J. Strand (1992), 'Willingness to Pay for Environmental Goods in Norway: A Contingent Valuation Study with Real Payment', *Environmental and Resource Economics* **2**, 91–106.

Sinden, J. A. (1988), 'Empirical Tests of Hypothetical Biases in Consumers' Surplus Surveys', *Australian Journal of Agricultural Economics* (August & December) **32**(2&3), 98–112.

Smith, V. K. and C. Mansfield (1998), 'Buying Time: Real and Hypothetical Offers', *Journal of Environmental Economics and Management* **36**, 209–224.

Spencer, M., S. Swallow and C. Miller (1998), 'Valuting Water Quality Monitoring: A Contingent Valuation Experiment Involving Hypothetical and Real Payments', *Agricultural and Resource Economics Review* **27**, 28–42.